Data 583 Life Expectancy (WHO)

Justin Chan, Kenny Tong, Viji Rajagopalan

22 Mar, 2023

Data 583 Life Expectancy - Final Report (Life Expectancy Data)

1. Introduction and Hypotheses

Life expectancy has always been an area of interest for humanity. The key to long live has remained an intriguing topic to people for decades. The goal of this project is to study a dataset that contains information on life expectancy and identify some of the variables that significantly impact life expectancy.

The dataset chosen for the study has life expectancy data of 193 countries between 2000-2015, together with different predictive factors. Broadly speaking, predicting variables are categorized into 4 major areas : Immunization, Mortality, Economical, and Social, containing a total of 21 individual variables. Our hypothesis is that a subset of variables from this dataset would be able to explain and predict life expectancy with good accuracy (say $> 80\%$). The dataset has a mix of variable types – continuous and discrete. Within discrete type, some variables are ordered, and some are unordered categorical variables.

With such a mix and complexity of data, we also hypothesize that all variables will not share a simple linear relationship with the predictor variable and modelling of life expectancy will require a more complex model. We analyze and validate several statistical models throughout the report with the primary goal of identifying an adequate model for the dataset.

2. Dataset overview

2.1 Variables Summary and Categories

Life expectancy is the response variable in this dataset. This represents the mean life expectancy (in age) by specific country and year combination. Refer **Table 1** and **Table 2** below for the list of predictor variables and their categories.

The dataset contains 2563 missing values in various columns. To handle the NA values in the dataset, two simple strategies are used. Firstly, 12 countries that have more NA values within their records are removed from the dataset. Secondly, the remaining NA records are imputed by the respective column mean.

Two columns are removed from the original dataset. The 'Percentage expenditure' variable is removed from the entire assessment as the values present in this column are unclear. Another variable 'country' is also removed because we intend to focus on studying the life expectancy on a global basis.

Variable	Unit of Measurement/Data Category	Continuous vs Discrete	Variable	Unit of Measurement/Data Category	Continuous vs Discrete
Life Expectancy	Years Old (Age)	Continuous	**Total expenditure**	Percentage	Continuous
Country	Nominal Data	Discrete	**Percentage expenditure**	Percentage	Continuous
Year	Ordinal Data	Discrete	**GDP**	Currency (USD)	Continuous
Status	Nominal Data	Discrete	**Population**	Count	Discrete
Adult Mortality	Count Data	Discrete	**Income composition of resources**	Percentage	Continuous
Infant deaths	Count Data	Discrete	**Schooling**	Mean (Years)	Continuous
Under-five deaths	Count Data	Discrete	**Alcohol**	Litres	Continuous
Hepatitis B	Percentage	Continuous	**HIV/AIDS**	Percentage	Continuous
Measles	Count Data	Discrete	**BMI**	Average BMI	Continuous
Polio	Percentage	Continuous	**Thinness 1-19 years**	Percentage	Continuous
Diphtheria	Percentage	Continuous	**Thinness 5-9 years**	Percentage	Continuous

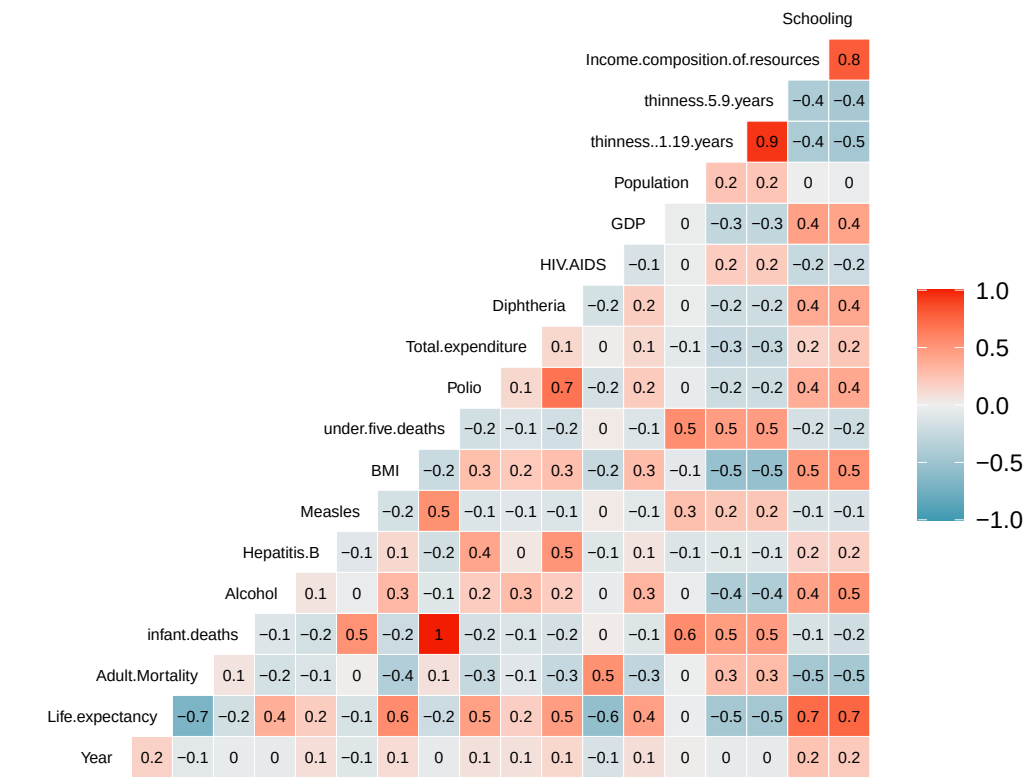
Table 1 : List of Predictor Variables

Following table shows how the above variables are grouped based on the 4 categories.

Data Categories	Variables
Economical Data	Total expenditure, Percentage expenditure, GDP, Income composition of resources
Social Data	Country, Status, Population, Schooling, Alcohol, BMI, Thinness 1-19 years, Thinness 5-9 years, BMI
Mortality Data	Adult Mortality, Infant deaths, Under-five deaths
Immunization Data	Hepatitis B, Measles, Polio, HIV/AIDS, Diphtheria

Table 2 : Variable Categories

The resulting dataset are then studied closely to understand their correlation effects with the response variable life expectancy. Following plot **Correlation Matrix Plot** is a correlation matrix on all the variables in the dataset.



It can be noted that the response variable life expectancy is highly correlated with income composition, schooling and adult mortality variables and has a correlation value of 0.7, 0.7 and −0.7 respectively. Life expectancy is moderately correlated with variables BMI, Polio, Diphtheria, HIV.AIDS and thinness variables and has a correlation value of 0.6, 0.5, 0.5, −0.6, −0.5.

2.2 Initial analysis using linear regression

Life expectancy is a continuous variable and the first choice is building a linear regression model which is simple and interpretable. A BIC backward step model variable selection method is also applied on the full model to arrive at a parsimonious model containing only significant predictor variables. Following table **Table 3** provides a summary of the two models.

Models	No. of Variables	Adj R-squared Score
Original Model	20	0.8299
Reduced Model	12	0.8296

Table 3 :Original vs Reduced Models

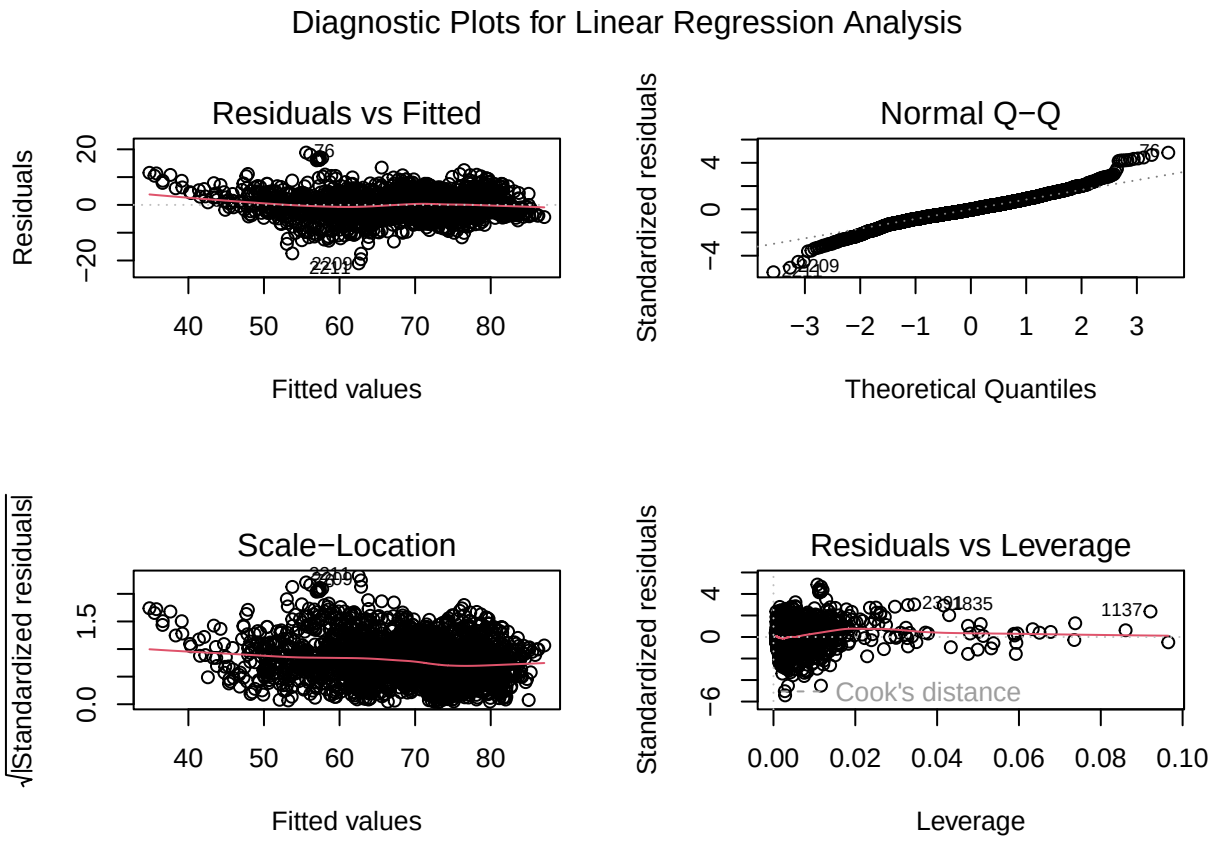
The number of independent variables is now effectively reduced to 12 and the reduced model contains the following 12 variables : Adult.Mortality + infant.deaths + Hepatitis.B + BMI + under.five.deaths + Polio + Diphtheria + HIV.AIDS + GDP + thinness..1.19.years + Income.composition.of.resources + Schooling. The adjusted R-squared score is well kept at nearly the same level as in the original model. The reduced model is able to explain more than 82% of variation in the response variable and its performance is above the anticipated 80%.

To conclude, the reduced model from this step is selected as the first model for the dataset. The dataset now has 2778 records and 12 columns. It is used for further evaluation from a linear model stand point and will be referred to as linear model in the remainder of the discussion.

3. Regression Analysis

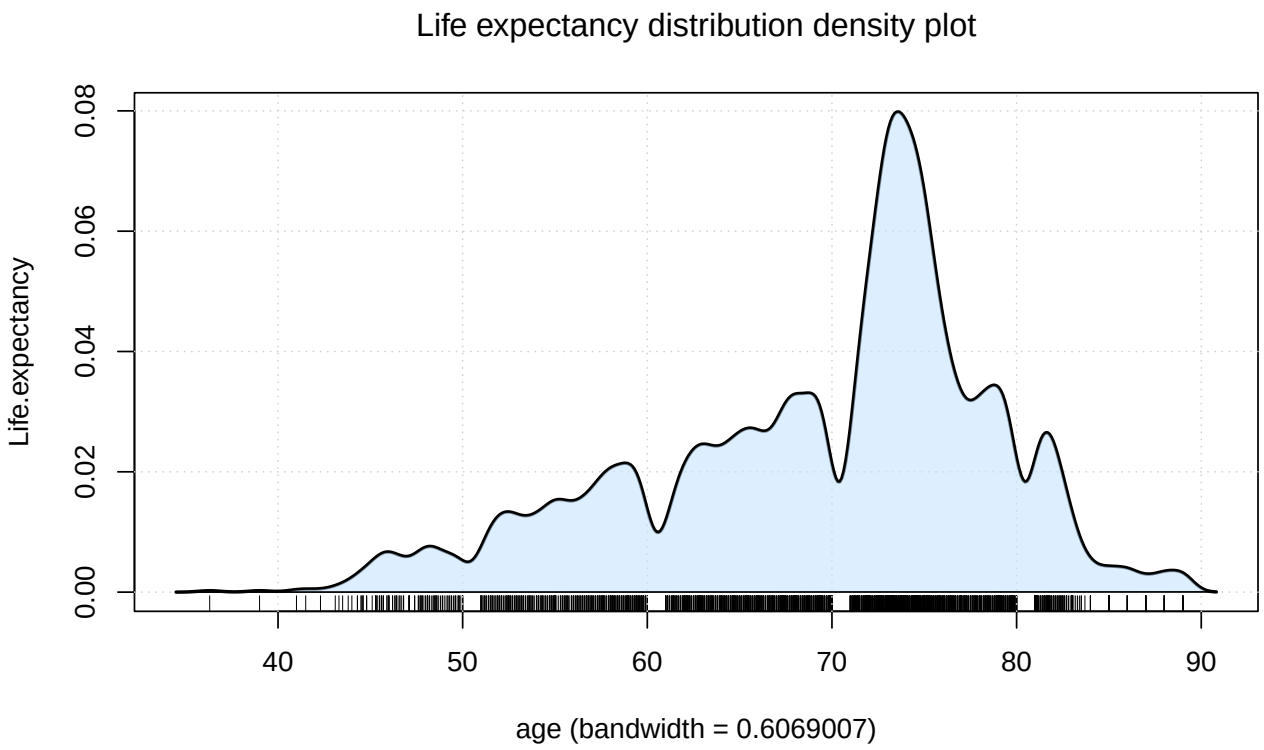
3.1 Linear model and diagnostics

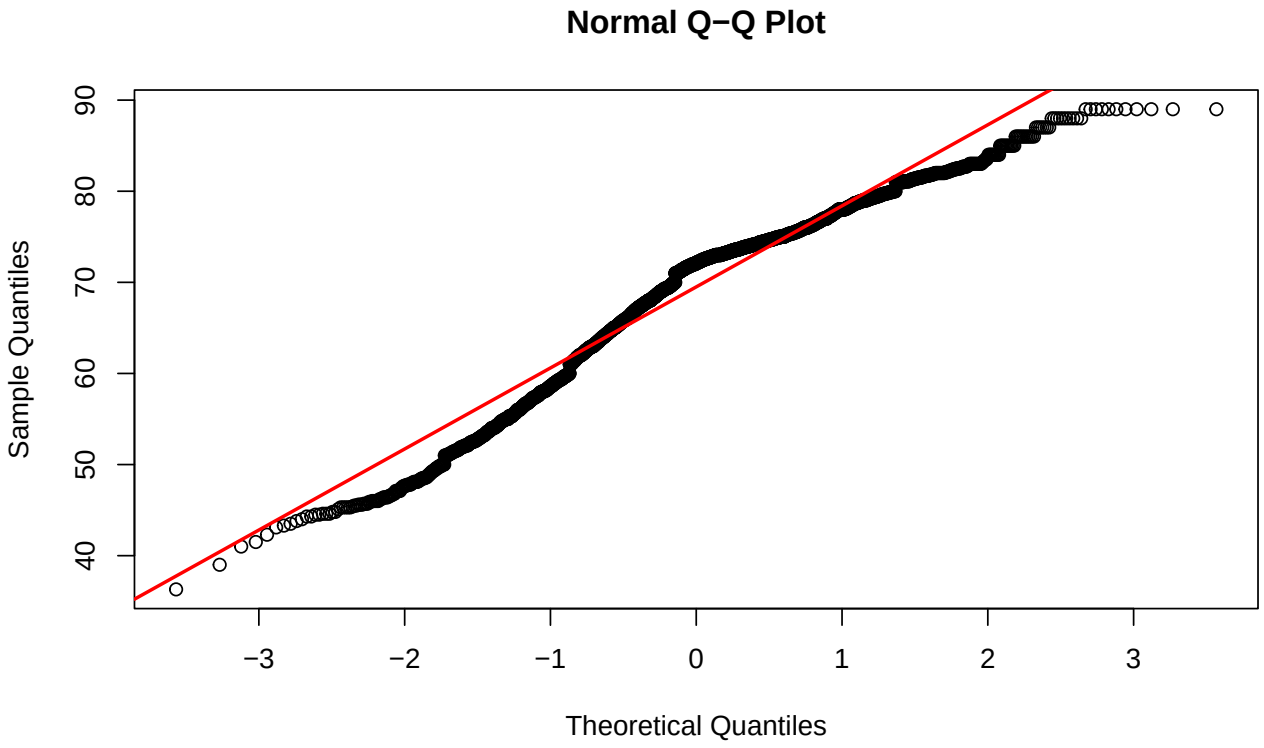
The linear model shows that we are able to explain approximately 82% of variability of our response variable using the selected predictor variables. The next step is to look at the error diagnostics from the model.



From the above diagnostic plots, the **Residual vs Fitted** and **Scale-Location** plots show a horizontal red line which indicate that the variance of the residual is same for different values of predictor variables. The **Residuals vs Leverage** plot shows some points with high residues and as the points are within cook's distance of 0.5, no records need to be eliminated. The **Normal Q-Q plot** suggests that the model is heavy tailed and the data on both ends of the quantiles do not fit on a straight line. This is an indication that the current linear regression based model is not fitting the data well and that the response variable is not normally distributed. Based on this, we undertake additional testing to validate if the model is adequate and valid.

a. Life expectancy variable distribution





From the **Life expectancy distribution density plot**, it can be seen that the mean of the distribution isn't symmetrical and the mean isn't centered at 0, indicating the response variable life expectancy is not normally distributed. From the **Normal QQ plot** of the regression model, it can be observed that there is a distinct curve in the middle of plot rather than a straight line, this indicates that there could be a bimodal distribution to response variable. These observations necessitate validation of distribution of the response variable.

b. Normal distribution test for response variable Shapiro-Wilk test is a statistical test for normality and a p-value that is very small and is less than 0.05 proves the variable in consideration is not normally distributed. Shapiro-Wilk test on the response variable life expectancy resulted in a p-value of $< 2.2e\text{-}16$.

This proves the response variable is not normally distributed and additionally, hypothesis tests for validating correct specification of parametric MLR models are conducted to identify if the selected linear model specification is valid for the given dataset.

c. Parametric model specification test Ramsey's RESET test is a test conducted to validate the correctness of the functional form. A p-value that is very small and is less than 0.05 rejects that the functional form is correctly specified. RESET test is conducted on the linear model and the resulting p-value is $< 2.2e-16$.

Based on the test, the linear model is rejected as the correct functional form for modelling the underlying data.

d. Consistent nonparametric inference The consistent nonparametric inference test is a hypothesis test for correct specification of parametric MLR models. This allows to estimate if the functional for given parameter estimates is reasonable when compared. A p-value that is very small and is less than 0.05 rejects that the functional form for given parameter estimates is reasonable.

As noted, the p-value for linear model is $< 2.22e-16$ and this output suggests that the linear model is rejected.

All the diagnostic tests indicate that linear regression is not an appropriate model for the given data. This proves one of the hypothesis of our project that a simple linear model may not be adequate in explaining the variability in the response variable life expectancy.

3.2 Parametric regression models and relative assessments

As the linear model is not adequate, we move on to other parametric regression models that do not assume normal distribution and known to perform well on complex and mixed data types. There are several models that could be used in the assessments and we selected LASSO and Neural Net with linear activation function for the given dataset.

The following **Table 4** shows variables that are selected for rest of the modeling based on correlation of the variables with the response variable, our knowledge on the domain. The selection process also focuses on prioritizing continuous variables. Here is a summary of the variable selection and comments describing the reason for removal of the variable. P.S: The variables not selected will be dealt with in subsequent runs to understand their significance and relationship with response variable.

Column Name	**Type**	**Reason of Removal**
Country	(Discrete)	Build models for all countries
Year	(Discrete)	Ordinal type data and based on domain knowledge, not prioritized
Status	(Discrete)	Nominal type data
Adult Mortality	(Discrete)	NA, Selected in first run
Infant deaths	(Discrete)	NA, Selected in first run
Under-five deaths	(Discrete)	NA, Selected in first run
Hepatitis B	(Continuous)	NA, Selected in first run
Measles	(Discrete)	Discrete type data and weak correlation with our predictor
Polio	(Continuous)	NA, Selected in first run

Column Name	Type	Reason of Removal
Diphtheria	(Continuous)	NA, Selected in first run
Total Expenditure	(Continuous)	Based on domain knowledge
Percentage Expenditure	(Continuous)	Based on domain knowledge
GDP	(Continuous)	NA, Selected in first run
Population	(Discrete)	No correlation with our predictor indicated by our correlation plot
Income composition of resources	(Continuous)	NA, Selected in first run
Schooling	(Continuous)	NA, Selected in first run
Alcohol	(Continuous)	Based on domain knowledge
HIV/AIDS	(Continuous)	NA, Selected in first run
BMI	(Continuous)	NA, Selected in first run
Thinness 1-19 years	(Continuous)	NA, Selected in first run
Thinness 5-9 years	(Continuous)	Range already covered in 1-19 Thinness 1-19 years

Table 4 : Variable selection for parametric and nonparametric models

First, the dataset is divided into a train and test datasets with an approximate 70 of the complete data using a random sampling process so long run performance of the models can be estimated. The train dataset has 2000 records and test dataset has 778 records in total.

Three types of models using linear regression (LM), LASSO and NeuralNet with linear activation function are built on the *train* dataset and the PRESS (Predicted Residual Error Sum of Squares) statistic is calculated using *test* dataset to identify the best performing model. Following **Table 5** summarizes and compares the PRESS statistics.

	LM	LASSO	NN
PRESS	15.93367	15.98972	22.43056

Table 5 : PRESS statistics for parametric models

Based on the output, it can be seen that LM and LASSO models perform better than NeuralNet model for this dataset. P.S: Linear model is used in these assessments for benchmarking purposes and not for actual use as the model is not valid.

Selecting LM and LASSO, the R^2 is also measured for the models. It can be seen from **Table 6** that LASSO performs nearly at the same level as the linear model (LM).

	LM	LASSO
R2	0.829139273435324	0.829061931452822

Table 6 : LM vs LASSO

Based on these assessments, LASSO is a viable model that can be considered for this dataset that has approximately 82 for predicted R^2 value and meets the performance goals expectations. LASSO does not have assumptions on the error distribution so residuals are not validated.

3.3 Nonparametric regression

Nonparametric regression is considered as another good option for the complex and mixed dataset that is of interest here due to proven flexibility and adaptability nature of these models. One important difference between nonparametric model and rest of the parametric models is that the entire data is used in the model training process.

The nonparametric regression is carried out with local linear estimator and cv.aic for automated bandwidth selection. This bandwidth selection method specifies expected Kullback-Leibler cross-validation (Hurvich, Simonoff, and Tsai (1998))and in general provides consistent estimates.

The output of the nonparametric regression model indicates an R^2 value of 87% approximately. This is the summary measure of in-sample fit for the model lies in the range of [0,1]. 1 denotes a perfect fit to the sample data and 0 indicates no fit. This is the counterpart to R^2 of linear model.

We acknowledget that the R^2 for LASSO model is calculated based on a train vs test setup and nonparametric regression R^2 is calculated on the complete dataset and a nonparamertic train-test fitting is identified as a future scope item that will be looked into. As the nonparametric model is a cross validated model and can provide long run performance, the LASSO and nonparametric model coeffients are compared and summarized in the **Table 7** table below.

	NPREG	**LASSO**
R2	0.8722143	0.829061931452822

Table 7 : Nonparametric vs LASSO based model comparison

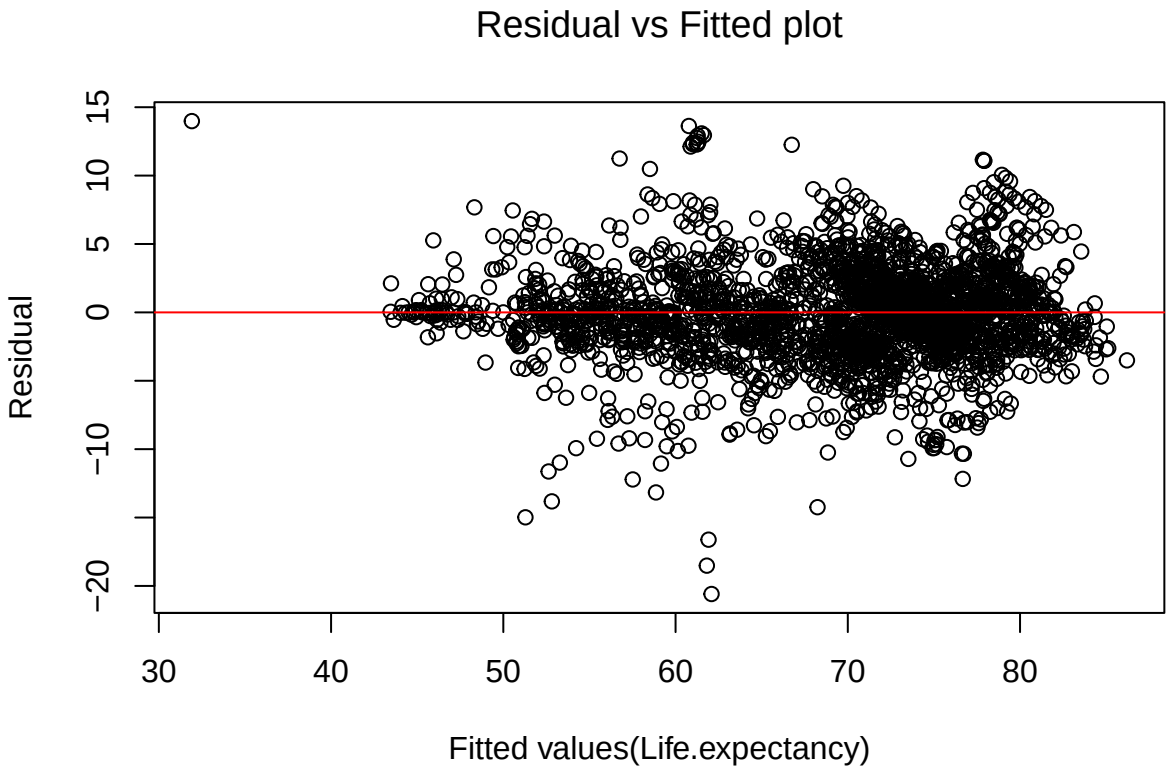
The nonparametric model has a higher R^2 of 87%(approx) when compared to the parametric model R^2 of 83%(approx). So, it is concluded that nonparametric model fits the given dataset better and is selected among the assessed models for use.

In order to arrive at a more parsimonious model, significance of the variables used in nonparametric regression is measured and per below **Nonparametric model variable significance** output all the variables are significant and will be retained in the model.

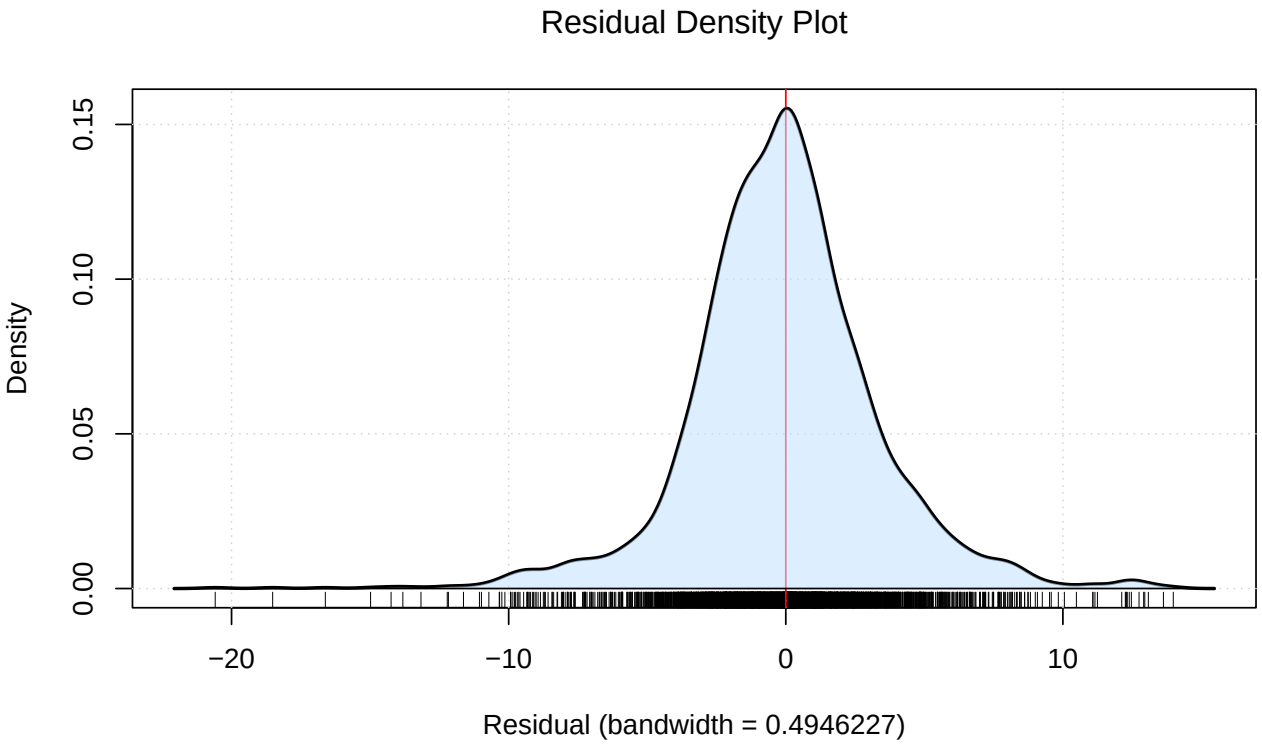
```
> npsigtest(model_np)

Kernel Regression Significance Test
Type I Test with IID Bootstrap (399 replications, Pivot = TRUE, joint = FALSE)
Explanatory variables tested for significance:
Adult.Mortality (1), infant.deaths (2), Hepatitis.B (3), BMI (4), under.five.deaths (5), Polio (6), Diphtheria (7), HIV.AIDS (8),
 GDP (9), thinness..1.19.years (10), Income.composition.of.resources (11), Schooling (12)

             Adult.Mortality infant.deaths
Bandwidth(s):       389457535       6733757
             Hepatitis.B      BMI under.five.deaths
Bandwidth(s):   225092161 79216285          95308072
               Polio Diphtheria HIV.AIDS
Bandwidth(s): 5825954   19248839 1.393258
                    GDP thinness..1.19.years
Bandwidth(s): 167351562078             37667667
             Income.composition.of.resources
Bandwidth(s):                         1071202
             Schooling
Bandwidth(s):  15165344

Individual Significance Tests
P Value:
Adult.Mortality                 < 2e-16 ***
infant.deaths                   < 2e-16 ***
Hepatitis.B                     0.047619 *
BMI                             < 2e-16 ***
under.five.deaths               < 2e-16 ***
Polio                           < 2e-16 ***
Diphtheria                      < 2e-16 ***
HIV.AIDS                        < 2e-16 ***
GDP                             < 2e-16 ***
thinness..1.19.years            < 2e-16 ***
Income.composition.of.resources < 2e-16 ***
Schooling                       < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: Nonparametric model variable significance

a. Error Diagnostics



Residual vs Fitted plot help in assessment of the goodness of fit for the nonparametric regression model. A well-fitted model typically show no pattern or structure in the residual plot, indicating that the error are randomly distributed around zero. Looking at the Residual plot vs fitted plots above, the majority of residual are randomly dispersed at around 0 suggesting that nonparametric regression model is acceptable and recommended for use.



Residual Density Plot is another method of evaluating models model fits, if the residuals follow a normal distribution, it is suggestion that the model is a good fit for the data. If the residuals deviate from a normal distribution, it is suggesting that the model may not be a good fit for our data and further analysis maybe needed. From our Residual Density Plot above, the distribution of our residual seems to form a bell-shape curve and centered at zero(indicated by the red line), suggesting that our residual is normally distributed and nonparametric regression model are a good fit for our data.

4. Conclusion

4.1 First set of models for predicting life expectancy

The two goals to evaluate and conclude during this project are a) to identify and arrive at a parsimonious model that can predict life expectancy to a level of > 80 with less variables than the orginal dataset b) to evaluate and show that a complex and mixed-type dataset is less likely to have all the predictor variables linearly related to the output response variable and thus likely will require a more flexible and complex model for better performance. On a side note, it is also an indicator that predicting life expectancy is complex.

Both these goals have been successfully evaluated during the course of the project and the conclusion is that life expectancy is by nature not a normally distributed response variable. Performing simple linear regression modelling on this complex data is not sufficient or valid. There are two models that are identified that can be used to predict life expectancy with R^2 of more than 80% aka the two models are able to explain variation in life expectancy to an extent of 80.

The nonparametric model has a higher R^2 of 87%(approx) when compared to the parametric model R^2 of 83%(approx). So, it is concluded that nonparametric model fits the given dataset better and is selected as the best one among all assessed models for use of predicting life expectancy.

In order to arrive at a more parsimonious model, significance of the variables used in nonparametric regression is measured and all the 12 variables are significant and recommended to be retained in the model. The variables selected in predicting life expectancy by the recommended models are: 'Adult.Mortality','infant.deaths','under.five.deaths','Hepatitis.B', 'BMI','Polio','Diphtheria','HIV.AIDS','thinness..1.19.years', 'Income.composition.of.resources','Schooling','GDP'.

4.2 Future improvements

Based on the knowledge of the domain and analysis from the data, different models and statistical tests are performed in the given time frame of this project to predict life expectancy. There are more opportunities for improvements we discovered during the course of this project and the team suggests the following for future exploration, studies and implementation to see if an even better-performing model can be attained.

1. Currently, variables are used "as-is" from the dataset. Some techniques for variable encoding/transformation and more flexible models like GAMs can be explored.
2. Performing nonparametric model analysis consumed substantial amount of computing resources. The results achieved so far is generally sufficient for measuring long run performance. While resources and time allow in the future, consider fine-tuning this model by enforcing dataset splitting into training and testing sets, which can provide a better account on the predictive performance on unseen data.
3. According to the earlier multicollinearity studies (Refer **Appendix**), correlation is found between the variables infant.deaths and under.five.deaths. It is understood that such correlation may cause undesirable effect on model accuracy, fitting and interpretation. To resolve this issue, it is suggested to remove one of the correlated variables with the higher VIF (Variable Inflation Factor) score.
4. Currently in the analysis, data is imputed for NA values (rather than removing the records with NA values) has been deployed to retain more information and the value generated by other valid data columns of the record. Although data imputation is a common industry practice, different null data handling techniques (apart from data imputation using mean) can be investigated in future and compared with current results.

4.3 Interesting challenges

Every project deals with an interesting amount of challenges and there are some interesting challenges encountered during the course of this project as well. The team needed to pivot constantly in order to efficiently and confidently come up with the first suitable model for predicting life expectancy.

1. Running npreg algorithm using R on this dataset is extremely time-consuming. It took 30 hours in a notebook computer. This undesirable situation has seriously constrained our flexibility in fine-tuning and re-running the model with different model settings such as different variables selection, train-test assessments because time is limited to compare different sets and select optimal fit.
2. Similarly, running model significance took more than 30 hours. This has caused similar consequence as the previous point 1.
3. As mentioned in the earlier analysis, the distribution is not normal and this limited the application of several simple models that have a strong assumption on the underlying distribution. The team also has very limited experience in bimodal or multimodal distributions so even though the response variable seems to be bimodal, no specific assessments or fitting is carried out.

5. Appendix

5.1 Checking for Multicollinearity

As Multicollinearity can potentially affect the accuracy of regression model, a correlation study is undertaken to understand and assess the situation. It is found that infant deaths and under.five.deaths are nearly 100% correlated. From the variable inflation factors (VIF) calculations as well, it is evident that the two variables have a high inflation factor of 166.764248 and 166.967503 respectively. Variable under.five.deaths need to be removed from the dataset and the models need to be assessed for their performance. LASSO model can handle multicollinearity automatically so there is no change in performance expected for the model.